Recognizing Images of Eating Disorders with Deep Learning

S. N. Counts[1*], L. Alkulaib[1*], J-L. Manning[1], R. Pless[1], J. Harnett[1], H. Xuan[1], R. Begtrup[2], D. A. Broniatowski[1]

George Washington University[1]

Children's National Health System, Washington, D.C.[2]

Eating disorders (ED) are pervasive and do not discriminate based on race, religion, gender, or socioeconomic status. Comorbidities include anxiety, depression, substance abuse, self-injurious behaviors, and history of trauma. ED are often a lifelong struggle, with approximately ⅔ of patients never achieving a full and sustained remission.

Exposure to media expressing "the thin ideal" can be triggering to individuals with ED as well as those at risk for developing them. Social media is rife with these triggers. Concurrent with the rise of social media, individuals with ED have created communities in which they support one another in the dangerous pursuit of this illness' goal: to be "thin enough." Websites promoting anorexia (pro-ana) and ED as lifestyle choices valorize acting on ED symptoms. Such sites teach those suffering or at risk from ED how to act on the illness and support them in doing so, putting them at risk for severe health complications.

The impact of images in this community far exceeds that of other communities surrounding physical and mental health issues. Therefore, it is essential that clinicians and family members be able to identify websites containing images associated with the promotion of ED to prevent exposure to these triggers. This research aims to automatically detect such triggering material, with the ultimate goal of designing parental and clinical controls.

We report on a proof of concept, machine learning approach to identify pro-ana content, trained on example data from online social media searches. The training data was chosen to compare pro-ana content with other content similar in demographics and photographic style, composed of the hashtag-based categories #proana, #selfie, #ootd, and #greek.

We randomly chose 20% of these images as test data and train the Resnet Deep Learning neural network to classify the remaining images. On test data this gives 81% classification accuracy—a significant improvement over chance (25%). These proof of concept results suggest that it is feasible to automatically detect social media sources with triggering material, informing the creation of tools that can assist clinicians and family members to improve health outcomes.

We used the classifier to make a web application that assesses how pro-ana a social media user's content is. The tool, designed for clinicians, allows them to enter a social media username and then gives an analysis of that user's online presence, classifying its content. The tool also displays a hashtag similarity map showing trending hashtags closely related to #proana.